Accelerated evolution of conserved noncoding sequences in the human genome

Shyam Prabhakar[1,2*], James P. Noonan[1,2*], Svante Pääbo[3] and Edward M. Rubin[1,2+]

1. US DOE Joint Genome Institute, Walnut Creek, CA
2. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA
3. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

*These authors contributed equally to this work.

+ To whom correspondence should be addressed. Email: emrubin@lbl.gov.

One sentence summary:

Conserved noncoding sequences exhibiting human-specific accelerated evolution are identified and shown to be enriched near genes involved in neuronal cell adhesion, suggesting a *cis*-regulatory contribution to the rise of human-specific cognitive traits.

Online Abstract

Changes in gene regulation likely influenced the profound phenotypic divergence of humans from other mammals, but the extent of adaptive substitution in human regulatory sequences remains unknown. We identified 1,119 conserved noncoding sequences (CNSs) with a significant excess of human-specific substitutions. These accelerated elements were disproportionately found near genes involved in neuronal cell adhesion. To assess the uniqueness of human noncoding evolution, we examined CNSs accelerated in chimpanzee and mouse. Although we observed a similar general trend towards neuronal adhesion in chimpanzee, the accelerated CNSs themselves exhibited almost no overlap with human, raising the possibility of independent evolution towards different neuronal phenotypes in each species. CNSs accelerated in mouse showed no bias toward neuronal cell adhesion. Our results indicate that widespread *cis*-regulatory changes in human evolution may have contributed to uniquely human features of brain development and function.

The distinctively human traits that distinguish us from all other primates originated in human-specific DNA sequence changes. To investigate whether gene regulatory or other functional noncoding elements in the human genome bear the signature of accelerated evolution, we determined the occurrence of human-specific substitutions in 129,405 conserved noncoding sequences (CNSs) previously identified by multiple whole-genome sequence comparisons (1).

We developed a test statistic that evaluated the likelihood of observing the configuration of human-specific substitutions present in a given CNS. We assigned each CNS a human-acceleration P-value based on the probability of observing a configuration of equal or smaller likelihood under the null model of constrained evolution (1). We identified 1,119 elements (0.86%) with a significant excess of human-specific substitutions at $P \leq 0.005$, 73% more than we would expect to see by chance at this P-value threshold (Figure 1A).

To ascertain in an unbiased manner if accelerated CNSs disproportionately occur near genes with particular functions, we determined the closest neighboring RefSeq gene for all 129,405 CNSs, obtained the Gene Ontology (GO) annotations for each gene, and assigned those annotations to each CNS. We then sought to identify GO terms with a significant excess of accelerated CNSs. *P*-value thresholds were set to adjust for multiple testing (1).

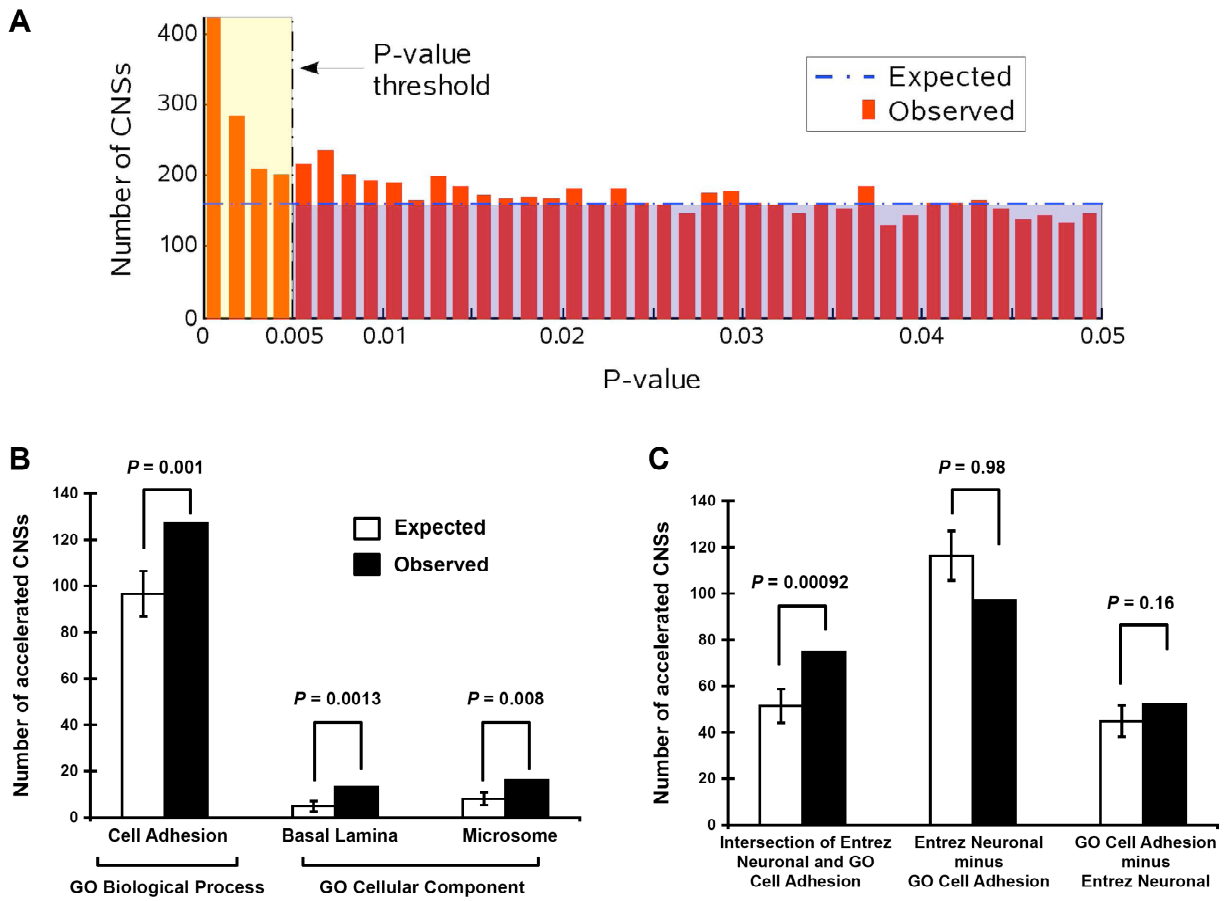

Figure 1. (A) Observed distribution of human-acceleration *P*-values in 129,405 CNSs versus the uniform distribution expected by random chance. **(B)** GO biological process and cellular component terms significantly enriched in accelerated CNSs. **(C)** Human-accelerated CNSs are disproportionately associated with genes functioning specifically in neuronal cell adhesion. There is a highly significant excess of accelerated CNSs that occur near genes with both GO cell adhesion and Entrez Gene neuronal annotations (*left*). However, the number of accelerated CNSs near genes with only Entrez Gene neuronal (*center*) or only GO cell adhesion annotation (*right*) is not significantly greater than expected by chance.

The GO cellular component term most significantly enriched in accelerated CNSs was basal lamina (Figure 1B). Of the 13 accelerated CNSs in this category, 10 were associated with the dystrophin-associated glycoprotein complex, disruptions of which cause muscle and neuronal diseases (2, 3). Cell adhesion was the only biological process displaying a significant excess of CNSs accelerated in human (Figure 1B). Many of the cell-adhesion accelerated CNSs were associated with genes involved in neuronal cell adhesion, such as cadherins and protocadherins, contactins, neuroligins, and classical neuronal cell adhesion molecules. To quantitatively evaluate this observation, we constructed a composite neuronal adhesion GO term by intersecting GO "cell adhesion" genes with genes annotated in the Entrez Gene database as having evidence of neuronal function. We found a highly significant excess of accelerated CNSs neighboring genes with both GO cell adhesion and Entrez Gene neuronal annotations (P = 0.00092, Fisher's exact test, one-sided; Figure 1C; Table S1D). However, when these overlapping accelerated CNSs were removed from the analysis, the number of accelerated CNSs with only GO cell adhesion or Entrez Gene neuronal function annotations was not significantly greater than expected. Thus, the strongest signal of human-specific noncoding sequence evolution we detected was an excess of accelerated CNSs near genes specifically involved in neuronal cell adhesion, rather than the more general categories of cell adhesion or neuronal function.

To determine if the pattern of noncoding sequence acceleration we observed in the human lineage was recapitulated in other lineages, we identified accelerated CNSs in chimpanzee and mouse (1). We observe 1,180 accelerated CNSs in chimpanzee, only 38 (3.2%) of which were also accelerated in human, indicating a general lack of overlap between human and chimpanzee accelerated CNSs (Table S1, A and B). While accelerated CNSs in chimpanzee showed little overlap with human, they were also significantly enriched near neuronal cell

adhesion genes (expected = 54, observed = 77, $P = 0.0017$; Table S1E). These results suggest independent accelerated evolution of neuronal cell adhesion functions in both the human and chimpanzee lineages. To determine if this was a general phenomenon among mammals, we examined the 5,058 CNSs accelerated in mouse and failed to detect any enrichment near genes involved in neuronal cell adhesion (expected = 234, observed = 207, $P = 0.97$; Table S1, C and F).

Our results suggest that the disproportionate association of accelerated CNSs with neuronal cell adhesion genes in human and chimpanzee reflects evolutionary processes specific to those lineages, rather than a general property of noncoding sequence acceleration in mammals. This observation is consistent with the rapid evolution of behavioral and cognitive traits seen in both humans and chimpanzees (4). Since the CNSs accelerated in the two lineages are largely disjoint, it is unlikely that the acceleration of neuronal adhesion CNSs in humans and chimpanzees results in the same neuronal phenotypes in the two species. These findings suggest that *cis*-regulatory and other noncoding changes may have contributed to the modifications in brain development and function that gave rise to uniquely human cognitive traits.

References and notes

1. Materials and methods are available as supporting material on Science online.

2. I. Dalkilic, L. M. Kunkel, *Curr. Opin. Genet. Dev*. **13**, 231 (2003).

3. M. P. Moizard *et al.*, *Eur*. *J*. *Hum*. *Genet*. **8**, 552 (2000).

4. A. Whiten *et al*., *Nature* **399**, 682 (1999).

5. We thank members of the Rubin lab for insightful discussions and support. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. J.P.N. was supported by NIH NRSA fellowship 1-F32-GM074367.

Supplemental Online Material

Materials and Methods

Table S1A-G

Supplemental online material

Materials and Methods

Identification of human-accelerated CNSs

CNS Filtering

We obtained whole-genome alignments as well as a genome-wide set of 186,675 human conserved regions identified in multiz 8-way genomic alignments by the phastCons program (1) as having a conservation score >= 400 from the UCSC Genome Browser (www.genome.ucsc.edu). These conserved regions were filtered for overlap with human mRNAs, human spliced ESTs, nonhuman mRNAs, retroposed genes or duplicated blocks annotated in the browser's self-chain track.

Neutral rate estimation

To quantify background noncoding (approximately neutral) evolutionary rates, which exhibit lineage- and locus-specific variation, we segmented the human genome into non-overlapping 1-Mb windows, appended marginal fragments shorter than 500 kb to the preceding window, excised all annotated known genes, retroposed genes, duplicated blocks and phastCons conserved regions (score >= 300), discarded windows aligned to < 50 kb in chimpanzee, mouse, rat or dog (insufficient data), and estimated substitution rates along each mammalian lineage in each window using fastDNAml (2). CNSs within retained windows were assigned background evolutionary rates and GC content by cubic-spline interpolation between window centers. CNSs with < 50 bp aligned in any of the five mammals were eliminated, and within-CNS substitution rates were estimated for the rest, assuming asymptoticity of the local background GC content.

Identification of human-specific substitutions

To identify CNSs within this set that contained an excess of human-specific substitutions, we empirically constructed a null model of human-lineage substitution probabilities in CNSs that accounted for four major sources of heterogeneity in CNS evolution: 1) variation in degree of constraint from position to position within a CNS, 2) variation in average constraint among CNSs, 3) lineage-specific variation in the local neutral rate of evolution (3) and 4) genome-wide relaxation of constraint in primates (4).

To account for rate variation among sites within a single CNS, we binned sites by their depth of conservation. We considered sites conserved in chimpanzee, rhesus macaque, mouse, rat, dog and chicken (Type 1, most constrained), sites conserved in chimpanzee, rhesus, mouse, rat and dog but substituted or absent in chicken (Type 2), and sites conserved only between chimpanzee and rhesus macaque (Type 3, least constrained). Rhesus orthologs were grafted from the multiz 17-way genome alignments into the 8-way alignments. We identified human-specific substitutions at Type1, Type 2 and Type 3 sites by parsimony as well as all Type 1, Type 2 and Type3 sites where human was identical to the other lineages. We filtered all sites for low sequence quality chimpanzee, rhesus, dog and chicken positions (Phred $Q < 30$) and for annotated human SNPs. We then generated counts of human-specific substitutions (K) and unsubstituted sites (N) of each type for each CNS.

Binning of CNSs by non-human constraint

To quantify evolutionary constraint in terms of sequence conservation, we defined the "constraint factor" C of a CNS or class of sites as the average ratio of its substitution rate to the local neutral rate. Thus, $C = 0$ implies extreme constraint, whereas $C = 1$ signifies neutral

evolution. Since C varies among lineages, and also among CNSs, we binned CNSs by their non-human constraint factor C_{NON}, which was determined from summed CNS and background rates over the chimpanzee, mouse, rat and dog lineages. We set an arbitrary threshold of $C_{NON} <= 0.4$ to eliminate phastCons predictions resulting from spurious alignments to low-complexity regions or human contamination in distant vertebrates, which yielded 129,405 CNSs for analysis of human-lineage acceleration.

Estimating human-specific substitution rate at each site type

The average human-lineage constraint factor $C^{T,n}{}_{HUM}$ at all sites in the genome within each two-dimensional category (defined by site type T and non-human constraint factor bin n) was estimated by maximum likelihood from counts of human-substituted $K^{T,n}(i)$ and conserved $N^{T,n}(i)$ sites in all CNSs i within the same bin n, and their associated human-lineage background neutral rates $R^{BG}{}_{HUM}(i)$ as: $C^{T,n}{}_{HUM} \approx \sum_i K^{T,n}(i) / \sum_i \left[R^{BG}{}_{HUM}(i)\, N^{T,n}(i)\right]$. Thus, the estimated human-specific substitution rate at a site of type T in CNS i, which lies in constraint-factor bin n is: $R^{T,n}{}_{HUM}(i) = R^{BG}{}_{HUM}(i) C^{T,n}{}_{HUM}$. To our knowledge, this is the first model of CNS evolution that accounts for variation of constraint among lineages and among sites within a CNS, as well as lineage- and locus-specific neutral rate variation.

Calculating acceleration P-values

The probability of specific human-lineage substitutions (for example, A->C) was calculated from the human-specific substitution rate $R^{T,n}{}_{HUM}(i)$ based on the HKY substitution model (5) parameterized by the local human GC-content and transition-transversion bias = 4.2, estimated from concatenated whole-genome CNSs using PAML (6). The negative log-probability of the observed human-specific substitution or conservation event at an individual site k is defined as

the site-specific human-lineage surprisal $s_{HUM}(k)$. If we assume that each CNS site evolves independently in the human lineage, the aggregate surprisal $S_{HUM} = \sum_k s_{HUM}(k)$ of a CNS constitutes an information-theoretic statistic summarizing the "surprisingness" of the observed configuration of human-lineage substitutions under the null model of sequence constraint, given the number of strong (G,C) and weak (A,T) Type 1, 2 and 3 positions within the CNS, the non-human constraint factor, the local human neutral rate and the local human GC-content. We calculated the CNS-specific probability distribution of S_{HUM} for each CNS under the null model of human evolution by convolving the distributions of the site-specific surprisals $s_{HUM}(k)$. The P-value of human-specific acceleration within a CNS is the probability of observing a surprisal greater than or equal to the actual value. By the definition of P-values, the expected number of CNSs in any P-value bin of width w is w*129,405 under the null model of constrained human-lineage evolution (w = 0.00125 in Figure 1). Thus, we expect only 647 human-accelerated CNSs at a P-value threshold of 0.005 (0.5%), though we observe 1,119 (0.86%).

Quantifying CpG effects

Since our null model of constrained noncoding evolution does not account for the high mutability of CpG dinucleotides, we performed a worst-case simulation of CpG noise by randomly introducing CpG to TpG transitions into human CNSs at a rate of 0.06 substitutions per CpG dinucleotide. This corresponds to the assumption that CpG to TpG transitions in CNSs occur at ~8 times the non-CpG rate (7), and that all such transitions in the human or chimpanzee lineage will be counted as human-specific substitutions. In reality, many of the chimpanzee-specific substitutions will be correctly identified as belonging to the chimpanzee lineage. We saw very little change in the number of accelerated CNSs detected, indicating that CpG effects are an unlikely explanation for the 472 extra accelerated elements we observe in the genome. This is

probably due to the fact that only ~1% of the dinucleotides in our CNS set are CpG.

Parallel studies in chimpanzee and mouse

Chimpanzee-specific CNS acceleration was identified exactly as described above, based on chimpanzee-specific Type 1, 2 and 3 substitutions. Since the aligned species set contains only two rodents (mouse and rat), the Type 3 parsimony category has no rodent equivalent. We therefore analyzed mouse-lineage CNS acceleration solely on the basis of Type 1 (conserved in human, chimpanzee, rat, dog and chicken) and Type 2 (conserved only in human, chimpanzee, rat and dog) positions. We verified that the human and chimpanzee GO term results described in the manuscript are robust even when Type 3 sites are excluded (data not shown). In all lineage-specific analyses, we started with the same initial set of human-based whole-genome alignments and human-based conserved noncoding elements, so as to maintain consistency in the input data.

Gene Ontology analyses

GO terms (http://www.godatabase.org/) of human RefSeq genes were augmented with those of their mouse and rat orthologs, as defined by Homologene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene). CNSs were assigned the human-mouse-rat biological process and cellular component GO terms of the closest human RefSeq gene (3' or 5' end). It is possible that a different gene might be closer to the CNS in the chimpanzee or mouse genomes. However, in analyzing chimpanzee- and mouse-specific CNS acceleration, we retained the human-based CNS annotations for consistency. Since 90% of the human and mouse genomes lie within syntenic blocks that are on average 6.9 Mb long, the human CNS-gene associations will in the majority of cases be preserved in mouse (8). By random chance 0.86% of human, 0.91% of chimpanzee and 4% of mouse CNSs associated with

a GO term are expected to be accelerated. Enrichment relative to this expectation was calculated using Fisher's exact test (one-tailed). Enriched parent GO terms of which all of the enrichment derives from a single daughter GO term were discarded. In order to ensure that the results reflected broad genomic trends rather locus-specific events, GO terms associated with < 10 accelerated CNSs (an arbitrarily chosen cutoff) were eliminated to minimize artifacts of multiple testing. Human genes were annotated as "neuronal" by searching for the keywords "neuron*," "neural*," "neurite," or "axon" on the Entrez Gene server (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene). Entrez Gene records are gene-specific, text-based annotations that are manually curated using data from the primary literature.

In all, we tested 3,594 biological process GO terms and 672 cellular component GO terms for association with accelerated CNSs. Given the hierarchical structure of the ontologies, and the fact that each gene in general has multiple GO-term associations, it is difficult to analytically correct GO-term enrichment *P*-values for the effect of multiple testing. We therefore estimated the effect of multiple testing by randomly labeling 1,119 human, 1,180 chimpanzee and 5,058 mouse CNSs as accelerated and testing each of the GO terms for enrichment in these CNSs. We performed 1,000 iterations of this randomization procedure, and estimated the *P*-value threshold at which only one GO term is expected to be significantly associated with accelerated CNSs by random chance. The resulting thresholds were $P = 0.004$ for human and chimpanzee biological process GO terms, $P = 0.0051$ for mouse biological process GO terms, $P = 0.024$ for human and chimpanzee cellular component GO terms, and $P = 0.030$ for mouse cellular component GO terms.

Correlation with recent positive selection in humans as revealed by polymorphism data

We examined the overlap between our accelerated CNSs and regions of the human genome showing evidence of recent selection in a genome-wide analysis of polymorphism data (9) but we saw no significant correlation between the datasets (data not shown). This is likely because the vast majority of accelerated CNSs we observe accumulated sequence changes throughout the course of human evolution following the divergence of the human and chimpanzee lineages. We are measuring acceleration over the entire ~6 million year course of human evolution since the divergence of humans and chimpanzees. SNP-based analyses of adaptive evolution can only detect selective sweeps occurring in the last 200,000 years of human evolution (10). This is $1/30^{th}$ of the time scale we are actually considering in our analysis.

Supplemental references

1. A. Siepel *et al*., *Genome Res*. **15**, 1034 (2005).
2. G. J. Olsen, H. Matsuda, R. Hagstrom, R. Overbeek, *Comput. Appl. Biosci*. **10**, 41 (1994).
3. C. Nusbaum *et al*., *Nature* **439**, 331 (2006).
4. G. V. Kryukov, S. Schmidt, S. Sunyaev, *Hum. Mol. Genet*. **14**, 2221 (2005).
5. M. Hasegawa, H. Kishino, T. Yano, *J. Mol. Evol*. **22**, 160 (1985).
6. Z. Yang, *CABIOS* **13**, 555 (1997).
7. J. Meunier *et al.*, *Proc Natl Acad Sci USA* **102,** 5471 (2005).
8. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
9. B. F. Voight *et al., PLoS Biol*. **4**, e72 (2006).
10. M. Przeworski, *Genetics* **160,** 1179 (2002).

Supplemental Tables

Table S1. **(A)** CNSs accelerated in the human lineage. N1, N2, N3: number of Type1, Type2 and Type3 positions in the CNS. K1, K2, K3: number of Type1, Type2 and Type3 substitutions in the CNS. Coordinates are based on the human May 2004 (hg17) genome assembly. **(B)** CNSs accelerated in the chimpanzee lineage (human coordinates). **(C)** CNSs accelerated in the mouse lineage (human coordinates). **(D)** Biological processes and cellular components significantly associated with CNSs accelerated in human. **(E)** Biological processes and cellular components significantly associated with CNSs accelerated in chimpanzee. **(F)** Biological processes and cellular components significantly associated with CNSs accelerated in mouse. **(G)** CNS and *P*-value thresholds for process and component associations of accelerated CNSs (adjustment for multiple testing).